BIOINFORMATICS ORIGINAL PAPER

Vol. 22 no. 2 2006, pages 220–225
doi:10.1093/bioinformatics/bti762

Genetics and population analysis

An efficient comprehensive search algorithm for tagSNP selection using linkage disequilibrium criteria

Zhaohui S. Qin[1,*], Shyam Gopalakrishnan[1,2] and Gonçalo R. Abecasis[1]

[1]Center for Statistical Genetics, Department of Biostatistics, School of Public Health, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029, USA and [2]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109-2122, USA

Received on August 6, 2005; revised on October 9, 2005; accepted on November 2, 2005
Advance Access publication November 3, 2005
Associate Editor: Martin Bishop

ABSTRACT

Motivation: Selecting SNP markers for genome-wide association studies is an important and challenging task. The goal is to minimize the number of markers selected for genotyping in a particular platform and therefore reduce genotyping cost while simultaneously maximizing the information content provided by selected markers.

Results: We devised an improved algorithm for tagSNP selection using the pairwise r^2 criterion. We first break down large marker sets into disjoint pieces, where more exhaustive searches can replace the greedy algorithm for tagSNP selection. These exhaustive searches lead to smaller tagSNP sets being generated. In addition, our method evaluates multiple solutions that are equivalent according to the linkage disequilibrium criteria to accommodate additional constraints. Its performance was assessed using HapMap data.

Availability: A computer program named FESTA has been developed based on this algorithm. The program is freely available and can be downloaded at http://www.sph.umich.edu/csg/qin/FESTA/

Contact: qin@umich.edu

Supplementary information: http://www.sph.umich.edu/csg/qin/FESTA/

INTRODUCTION

With the rapid improvement of high-throughput genotyping technologies, genome-wide association studies are emerging as a promising approach to detect genetic variants that contribute to human diseases. Initially, genome-wide association studies will focus on single nucleotide polymorphisms (SNPs) because of their high abundance in the human genome, their low mutation rates and their accessibility to high-throughput genotyping (Collins *et al*., 1997). There are more than 10 million verified SNPs in dbSNP (build 124) (Sachidanandam *et al*., 2001), but typing all available SNP markers is inefficient and not necessary since many will provide redundant information due to linkage disequilibrium (LD). A better strategy is to select a subset of representative SNPs (tagging SNPs or tagSNPs) and to remove the rest from consideration (Johnson *et al*., 2001; Cardon and Abecasis, 2003). The objective is to have little information overlap among the selected SNPs while retaining much of the signal contained in the original set.

The selection of tagSNPs has become a very active research topic and many strategies have been proposed (Patil *et al*., 2001; Zhang *et al*., 2002; Gabriel *et al*., 2002; Johnson *et al*., 2001; Meng *et al*., 2003; Sebastiani *et al*., 2003; Avi-Itzhak *et al*., 2003; Ke and Cardon, 2003; Goldstein *et al*., 2003; Stram, 2003; Hampe *et al*., 2003; Chapman *et al*., 2003; Lin and Altman, 2004; Halldórsson *et al*., 2004; Rinaldo *et al*., 2005). Recently, Zhang and Jin (2003) and Carlson *et al*. (2004) introduced methods based on the LD measure r^2. These methods search for a small set of SNPs that are in strong LD (measured through pairwise r^2) with other SNPs that are not selected for genotyping. Pairwise r^2 is an attractive criterion for tagSNP selection since it is closely related to statistical power for case–control association studies, where a directly associated SNP is replaced with an indirectly associated tagSNP (Pritchard and Przeworski, 2001).

In this manuscript, we describe efficient algorithms for tagSNP selection based on pairwise LD measure r^2. The algorithms were implemented in a computer program named FESTA (fragmented exhaustive search for tagging SNPs). Essentially, we replace a greedy search, where markers are added sequentially to the tagSNP set, with an exhaustive search where all marker combinations are evaluated. To achieve this, we arrange the genome into precincts of markers in high LD, such that markers in different precincts show only low pairwise disequilibrium. TagSNP selection can then be performed within each precinct independently, greatly reducing computation cost. In most settings, our method is guaranteed to find the optimal tagSNP set(s) defined by the r^2 criterion. For a small proportion of precincts where exhaustive search is computationally too expensive to carry out, an efficient greedy-exhaustive hybrid search algorithm is described. Using data from the HapMap project (The International HapMap Consortium, 2003), we show that the majority of these precincts contain relatively small numbers of SNPs, especially when a stringent r^2 criterion is used. Our algorithm readily identifies equivalent tagSNP sets, so that additional selection criteria can be incorporated. Other useful extensions are also discussed in this manuscript, such as the inclusion/exclusion of certain SNPs and double coverage, which can increase robustness of tagSNP sets against sporadic genotyping failures or errors.

METHODS

Consider a set $\mathbb{S}$ which contains M bi-allelic SNP markers $a_1, a_2, \ldots, a_M$. Further assume that all these markers have minor allele frequency (MAF) above a certain threshold (0.05 was used in this study). First, two-SNP haplotype frequencies were estimated (Hill, 1974), and then the pairwise LD measure r^2 (also referred to as 'Δ^2') (Devlin and Risch, 1995) was

[*]To whom correspondence should be addressed.

calculated for each pair of markers using the inferred haplotype frequencies (Hill and Robertson, 1968). Two markers a_i and a_j are said to be in strong LD if the r^2 between them is greater than a pre-specified threshold value r_0, denoted as $r^2(a_i, a_j) \geq r_0$ ($r_0 = 0.5$ or 0.8 in in this study). Both are considered tagSNPs for each other; in that a_i can be used as a surrogate for a_j, or vice versa.

Our aim is to a find tagSNP set, denoted by T, a subset of $\mathbb{S}$ such that $\forall a_i \in \mathbb{S} \backslash T$, $\exists a_j \in T$ that satisfies $r^2(a_i, a_j) \geq r_0$. In our presentation, we introduce two intermediate SNP sets, P and Q. P is called the *candidate set* which contains all the markers that are eligible to be chosen as tagSNPs and Q is named the *target set* which contains all the markers that are yet to be tagged, i.e. no marker in Q is in LD with any tagSNP in T. For each marker a_m in P, let $C(a_m) = \{a : a \in Q \text{ and } r^2(a, a_m) \geq r_0\}$ represent the subset of Q which contains markers that are in strong LD with a_m, and let $|C(a_m)|$ be the number of the elements in the set $C(a_m)$. Typically, the candidate set P is the complement of the tagSNP set T, $P = \mathbb{S} \backslash T$ and $P = Q$. One exception occurs when some SNPs are excluded as tagSNPs because they cannot be easily genotyped, but they still should be tagged by other markers if possible. In this case, the candidate set is a subset of target set. We describe several different algorithms for updating P, Q and T starting with a greedy approach (Carlson *et al*., 2004). We then outline successive refinements and extensions of a partition and exhaustive search algorithm, designed to handle various scenarios encountered when planning association studies.

Greedy approach

The detailed algorithm is as follows (Carlson *et al*., 2004).

Algorithm 1 (greedy approach):

```
(1) Set $T = \emptyset$ and $P = Q = \mathbb{S}$;
(2) For each marker $a_m$ in $P$, calculate $|C(a_m)|$;
(3) For every marker $a_m$ where $|C(a_m)| = 0$, add $a_m$ into $T$, and remove it
    from $Q$;
(4) Find the marker in $P$ that has the highest $|C(a_m)|$ value, denoted as $a_{\max}$,
    and add $a_{\max}$ into $T$, removing it and all connected SNPs, i.e. $C(a_m)$
    from $Q$;
(5) Repeat Steps 2–4 until $Q = \emptyset$.
```

In Step 4, by removing associated markers from consideration, the coverage overlap among tagSNPs is greatly reduced. Although it is simple to implement, the greedy procedure may miss more efficient solutions. Figure 1 gives a simple example, where markers A and B each tag half of all markers and together can tag all the markers. However, marker C is connected to more than half of all markers, and it is the first marker selected by the greedy algorithm. In this example, the greedy algorithm produced a set with three tagSNPs, despite the fact that the optimal solution contains only A and B.

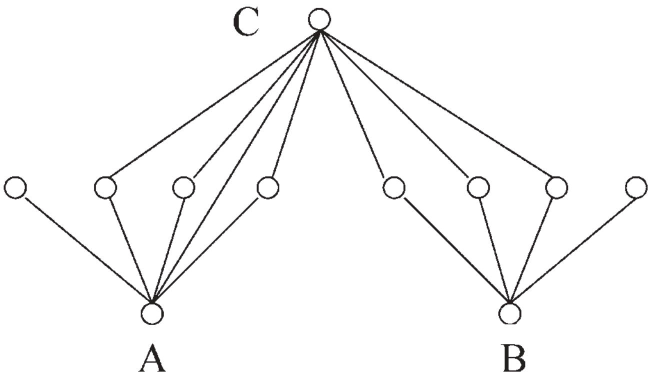

Fig. 1. An example when the greedy approach does not provide the smallest tagSNP set.

FESTA

An exhaustive search guarantees the minimum tagSNP set. Therefore, theoretically, the exhaustive search solves the tagSNP selection problem. But in practice, genome-wide tagSNP selection requires consideration of hundreds of thousands of SNP markers. For problem of this scale, exhaustive searches cannot be directly applied due to prohibitive computation costs.

Since appreciable LD only occurs within clusters of nearby markers along chromosomes, a practical solution is to first decompose the set of markers into disjoint precincts, such that markers in different precincts are never in strong LD. Then, selecting tagSNPs using the r^2 criterion in the whole set is equivalent to selecting tagSNPs in each precinct and then combining all the tagSNPs together. Here the concept of precinct is defined based on pairwise LD measure. It is therefore closely related to haplotype blocks (Reich *et al*., 2001; Patil *et al*., 2001; Daly *et al*., 2001; Jeffreys *et al*., 2001; Gabriel *et al*., 2002; Dawson *et al*., 2002), which are regions where historical recombination events are rare. The main difference is that the precincts of markers in high LD are determined purely on genetic distance. Unlike haplotype block, markers within each precinct may not be consecutive markers sitting next to each other.

Partitioning the markers into precincts can be achieved using standard algorithms in graph theory. We applied the Breadth First Search (BFS) algorithm (Cormen *et al*., 2001). Starting from any node (a marker) in a new precinct, this algorithm adds all neighboring nodes (markers in LD) and all neighbors of the newly added nodes to the precinct, until there are no neighbors to be added to the precinct. This process is restarted from different nodes until all the nodes are assigned a precinct.

After the partitioning step, we perform the tagSNP selection within each precinct. Starting with $K = 1$, all K-marker combinations are searched to see if they cover all markers within this precinct. If not, K is increased by one and the search is repeated until a tagSNP set is found or a pre-specified search limit is reached.

When evaluating all K-marker combinations, the computation cost required for an exhaustive search might be too great in some precincts. In such cases, we propose a hybrid solution which reduces the computation cost and retains a good chance of finding optimal tagSNP sets. For each precinct i with N_i markers (Here on, all parameters with subscript i indicate parameters within the i-th precincts, such as K_i, J_i, P_i, Q_i, T_i and N_i.), we decide whether an exhaustive search is feasible by comparing the computation cost required for evaluating all K-marker combinations within a precinct, $\binom{N_i}{K}$, with a computation cost limit L specified a priori, determined based on available computing resources. Larger limits allow a more comprehensive search, which may result in fewer tagSNPs being selected, but require additional computational effort. In this study, we set this limit at 1 million. When this limit is exceeded, we apply the following hybrid algorithm. Specify K_i^* such that it is the largest K possible that satisfies $\binom{N_i}{K} \leq L_0$, where L_0 is a pre-specified computation cost limit (less than L, set at 10 000 in studies conducted here). Subsequently, for each K_i^*-marker combinations, denoted as $\{a_1, \ldots, a_{K_i^*}\}$, assume that these markers have already been selected, remove a_m together with all the markers in $C(a_m)$ from candidate set P_i and target set Q_i, $m = 1, \ldots, K_i^*$, i.e. $P_i = Q_i = \mathbb{S}_i \backslash \cup_{m=1}^{K_i^*} (\{a_m\} \cup C(a_m))$ then apply the greedy approach to identify a subset of P_i that is able to cover Q_i, which contains the remaining untagged markers. The tagSNP set obtained in the reduced set plus the previous K_i^* markers together form a complete tagSNP set for the i-th precinct. The detailed algorithm is as follows:

Algorithm 2 (FESTA: greedy-exhaustive hybrid search):

```
(1) Apply the Breadth First Search to decompose the entire set of markers
    into precincts such that high LD can only be observed within precincts.
    $\mathbb{S} = \cup_{i=1}^{n} \mathbb{S}_i$, and $\mathbb{S}_i \cap \mathbb{S}_j = \emptyset$ for all $i \neq j$;
```

(2) Within each precinct $\mathbb{S}_i$, set $K = 1$,

   (a) If $\binom{N_i}{K} \geq L$, move to b, otherwise, conduct an exhaustive search over all possible K-marker combinations. Both the candidate set P_i and the target set Q_i is $\mathbb{S}_i$. If no combination of K SNPs can cover the entire precinct, set $K = K + 1$, and repeat this step;

   (b) Find K_i^* such that $\binom{N_i}{K_i^*} \leq L_0$ and $\binom{N_i}{K_i^* + 1} > L_0$. For every K_i^*-marker combination in $\mathbb{S}_i$, denoted as $\{a_1, \ldots, a_{K_i^*}\}$, let $T_i = \cup_{m=1}^{K_i^*} \{a_m\}$, $P_i = Q_i = \mathbb{S}_i \backslash \cup_{m=1}^{K_i^*} \quad (\{a_m\} \cup C(a_m))$, and apply the greedy approach to identify a subset of P_i that is able to cover the remaining untagged markers Q_i. Among all the resulting tagSNP sets, we choose the smallest.

(3) Record all minimum tagSNP sets that cover the precinct. These form the complete minimum tagSNP sets $\{T_i^j : j = 1, \ldots, J_i\}$, where J_i is the total number of such minimum tagSNP sets.

(4) Any combination of tagSNP sets identified from all disjoint precincts forms a tagSNP set for the whole set $\mathbb{S}$. Suppose the size of the minimum tagSNP set(s) in each precinct is K_i, then the overall size of such minimum tagSNP sets is $\sum_{i=1}^n K_i$, and the total number of such minimum tagSNP sets is $\prod_{i=1}^n J_i$.

FESTA executes either a ‘pure exhaustive search’ or a ‘greedy-exhaustive hybrid search’ in each precinct depending on the computational cost. Exhaustive search is first attempted, and if the computation cost becomes too high, the hybrid algorithm is used as a fallback. Typically, only a small proportion of the precincts require the ‘greedy-exhaustive hybrid search’.

FESTA double coverage

So far, both the greedy approach and our FESTA algorithm focus on finding a tagSNP set such that each SNP is either a tagSNP itself or is in LD with at least one of the tagSNPs. This is a criterion aimed at minimizing the number of tagSNPs selected. In reality, random genotyping failure or genotyping error on these tagSNPs can result in loss of power to identify the true signal. To be more robust against such adverse events, we evaluated a more stringent criterion requiring that every untyped SNP be in LD with at least two tagSNPs.

Our FESTA algorithm can be extended to find tagSNP sets that will have double coverage on the SNP markers considered. As always, an exhaustive search is able to find such tagSNP sets when the marker set considered is not too large. When exhaustive search is not feasible, the same greedy-exhaustive hybrid search strategy can be applied. The detailed FESTA double coverage algorithm can be found in the Supplementary Online Material. Note that in practice, it may be useful to consider double coverage only for large precincts, where the cost of losing an SNP to genotyping failure might be higher.

Further tagSNP selection considerations

Mandatory tagSNP markers Our algorithm readily allows users to force certain SNP markers to be included in or excluded from the tagSNP set. There are several scenarios where such functionality is important. First, in candidate gene studies, previous knowledge may be available as to which SNPs are functionally important. These might include non-synonymous coding region SNPs (cSNPs) as well as SNPs located in regulatory regions. Second, in genome-wide studies, one might carry out multiple rounds of genotyping and tagSNP selection. In such cases, additional tagSNPs could be selected at each round to cover the markers not tagged by tagSNPs successfully genotyped in the previous round. In other settings, it may be useful to exclude certain SNPs from consideration as tags. For example, some SNP markers may be difficult to genotype using a particular platform.

When there are mandatory markers $t_1, t_2, \ldots, t_r$, to be included, put these markers into the tagSNP set T and remove them from the candidate set, e.g. P becomes $P \backslash \cup_{i=1}^r \{t_i\}$. The target set Q becomes $Q \backslash \cup_{i=1}^r (\{t_i\} \cup C(t_i))$. If there are SNPs $u_1, u_2, \ldots, u_s$ that need to be excluded from the tagSNP set, remove them from the candidate set P, the target set Q is unchanged.

Choosing between alternative solutions Within a densely typed SNP set, redundant tagSNPs are common, which results in multiple tagSNP sets of the same size. All of these sets are equal in the sense of minimizing the number of tagSNPs. In order to choose one set for genotyping, additional criteria can be entertained. Here are examples of such additional criteria:

(1) Maximize average r^2 between tagSNPs and untagged SNPs they represent;

(2) Maximize the lowest r^2 between tagSNPs and the untagged SNPs they connect to;

(3) Minimize the average r^2 among all pairs of tagSNPs within a precinct.

In Criteria 1 and 2, we try to identify tagSNP sets whose members have the strongest connections with those untagged SNPs. Since pairwise r^2 between disease locus and marker loci is closely related to statistical power of detecting association, this strategy should result in increased power on average and in the worst case, respectively. These criteria are recommended in regular association study designs. The purpose of using Criterion 3 is to find a tagSNP set whose members are as independent as possible which minimizes overlap between covered SNPs of different tagSNPs and potentially increases the chance of linking to untyped SNPs. This strategy is particularly useful if one suspects the actual disease locus is not among the marker loci genotyped. To evaluate the potential of uncovering the disease-causing mutations in association studies among tagSNP sets identified by the aforementioned criteria, we conducted some empirical evaluations, summarized in the Results section.

Other types of criteria may be of even greater interest in practice. For example, in many genotyping technologies, some SNPs are harder to genotype than others due to the characteristics of surrounding genome sequence. We can use this information to select tagSNPs that are likely to have a high success rate and to avoid SNPs that are prone to genotyping failure.

RESULTS

In order to illustrate our proposed piecewise exhaustive search strategy, compare it with the greedy approach and explore the various characteristics of the tagSNP sets selected by our method, we applied both methods to two sets of data, the entire Chromosome 2 and five ENCODE regions (ENr112, ENr131, ENr113, ENm010 and ENm013) genotyped by the HapMap project (release 16c, June 2005). All three populations: CEU (European), YRI (Yoruban) and JPT + CHB (Japanese and Chinese) were studied. The first is in the context of a genome-wide association study and the second is similar to the situation of a candidate region study.

Chromosome-wide tagging

We have applied the greedy algorithm and FESTA to Chromosome 2 using HapMap Phase 1 genotype data (release 16c, June 2005). Table 1 (r^2 threshold of 0.5) and Table S1 (r^2 threshold of 0.8) summarize the results. FESTA produces less tagSNPs compared with the greedy approach in all three populations. When compared across populations, the YRI samples have about twice the amount of tagSNPs as the CEU or the JPT + CHB samples. The JPT + CHB samples have slightly less tagSNPs identified than the CEU samples. With r^2 threshold 0.5, the percentages of tagSNPs identified by our new algorithm are 21.6% in CEU, 39.3% in YRI and 20.9% in JPT + CHB samples, respectively.

The size of the tagSNP set is optimal for precincts where the greedy approach indicates that one or two tagSNPs are enough to cover all the SNPs in it. Improvements over the greedy approach is

Table 1. Summary of Chromosome 2: size of disjoint precincts and number of SNPs and tagSNPs in each precinct

	CEU	YRI	JPT + CHB
No. of SNPs	64 801	69 630	57 810
$r^2 \geq 0.5$			
No. of precincts	11 786	24 752	10 248
No. of tagSNPs (Greedy)	14 384	27 804	12 454
No. of tagSNPs (FESTA)	13 983	27 379	12 108
No. of tagSNPs (FESTA, double cover)	23 644	41 668	20 644
$r^2 \geq 0.8$			
No. of precincts	23 426	41 079	20 178
No. of tagSNPs (Greedy)	24 300	41 729	21 044
No. of tagSNPs (FESTA)	24 176	41 664	20 963
No. of tagSNPs (FESTA, double cover)	35 824	54 101	31 463

Table 2. Distributions of the size of the tagSNP sets using the greedy approach and the FESTA algorithm (with r^2 threshold of 0.5)

	CEU		YRI		JPT+CHB	
	Greedy	FESTA	Greedy	FESTA	Greedy	FESTA
Singleton[a]	5241	5241	15 079	15 079	4416	4416
1	5172	5172	8096	8096	4660	4660
2	774	911	924	1070	634	770
3	318	278	355	291	312	250
4	144	99	127	100	113	90
5	59	42	73	53	60	30
6	27	18	36	28	16	15
7	17	17	21	10	14	6
8	16	4	10	8	11	6
9	11	1	6	3	4	4
10+	7	3	25	14	8	1
Total	14 384	13 983	27 804	27 379	12 454	12 108

[a]Singleton means the precinct only contains one SNP. In other words, singleton refers to an SNP marker that is not in LD (pairwise LD measure r^2 greater than a threshold) with any other SNP in the entire set. Such an SNP, by definition, is one of the tagSNPs.

only possible for the remaining precincts. In the CEU samples, there are 599 of such precincts, in which the greedy approach identified 2423 tagSNPs, and FESTA identified 2022, a 16.5% reduction. When the r^2 threshold is 0.8, 154 precincts require more than two tagSNPs, as identified by the greedy approach. Among them, the greedy approach and FESTA identified 526 and 402 tagSNPs, respectively. The reduction rate is 23.6%. All the detailed results are summarized in Table 2 (r^2 threshold of 0.5) and S1 (r^2 threshold of 0.8). When double coverage is required, 69.1 and 45.9% more tagSNPs are needed with r^2 thresholds of 0.5 and 0.8, respectively. Similar results were obtained from the YRI and JPT + CHB samples.

Among all the non-singleton precincts in the CEU samples (6545 for r^2 threshold of 0.5 and 10196 for r^2 threshold of 0.8), most require only a small number of tagSNPs, so that the exhaustive search can be applied directly. With r^2 threshold of 0.5, the greedy-exhaustive hybrid approach was required for only 98 precincts or 1.5% of all precincts (11 precincts (0.1%) with r^2 thershold of 0.8).

Densely typed region

A very dense SNP map was recently released by the HapMap project on the ENCODE regions. We used five such regions (ENr112, ENr131, ENr113, ENm010 and ENm013) to evaluate the performance of our algorithm. Each ENCODE regions is 500 kb in length, for the CEU samples, the average number of SNPs in these regions is 832 (ranges from 551 to 1126), corresponding to an SNP density about 1 SNP per 601 bps (1 SNP per 907 bps to 1 SNP per 444 bps for individual regions). The detailed results were summarized in Table 3. Detailed results for the YRI and JPT + CHB samples can be found in Supplementary Tables S2 and S3.

In this set of densely typed SNPs, using our method with r^2 threshold of 0.5, the average percentage of tagSNPs required to cover each of the five regions is 8.3% of all markers (ranges from 5.4 to 11.3%). For double coverage, on average, 76.7% more tagSNPs are required (ranges from 70.7 to 83.6%). With a more stringent r^2 threshold of 0.8, the average percentage of tagSNPs required increased to 16.6% of all markers (ranges from 11.4 to 24.1%). To double cover these regions, on average, 62.9% more tagSNPs are required (ranges from 56.9 to 71.6%). For those precincts where improvement over greedy search is possible, using FESTA, the reduction rate is 17.9 and 23.0% on average for the five ENCODE regions with r^2 thresholds of 0.5 and 0.8, respectively. Applying our method to YRI and JPT + CHB samples reveals similar trends (data not shown).

Additional TagSNPs for denser SNP map

With the rapid advance of genotyping technologies, progressively denser SNP maps will become available. As more refined association studies are carried out, it will be useful to select new tagSNPs to ‘fill holes’ in the initial sparse maps. With a good picking strategy for the first round of tagging, this staged approach should result in only a small-to-moderate increase in the total number of tagSNPs compared to a one-stage strategy.

To evaluate this strategy, we constructed an artificial sparse SNP map for each of the five ENCODE regions (using the CEU samples only). Specifically, we selected one in every five consecutive SNP markers. The density of this sparse map is about 1 SNP per 3kb, close to the density of the phase I HapMap. Then, three different tagSNP sets are identified using the three criteria described previously, denoted by T_i, $i = 1, 2, 3$. Finally, we applied our approach to the full ENCODE SNP set, using each of these tagSNP sets as a seed, so as to search for additional tagSNPs to cover the previously ‘hidden’ SNP markers. The effectiveness of these tagSNP sets is evaluated by comparing the number of new tagSNPs needed to cover the ‘newly found’ SNPs. In addition to the three criteria, we also compared three other tagSNP selection strategies: Z random SNPs, assume Z is the number of tagSNPs for the sparse map; a picket fence strategy with Z equally spaced SNPs (where we place equally spaced grid points along the interval and then select markers that are closest to these grid points); or using all original SNPs as tagSNPs. The results are summarized in Table 4 (r^2 threshold of 0.5) and Table S4 (r^2 threshold of 0.8) in the Supplementary Online Material. From there, one can see that when the r^2 threshold is 0.5, 14.4% more tagSNPs (range from 7.0 to 20.9%) are needed to fill holes in the original map and that number is only 5.4% (range from 3.8 to 7.0%) when r^2 threshold is 0.8. The three tagSNP sets require

Table 3. Summary of TagSNPs identified by the greedy approach, the FESTA and FESTA double coverage algorithms in the five ENCODE regions (CEU samples)

Region	ENr112	ENr131	ENr113	ENm010	ENm013
No. of SNPs	863	988	1061	539	708
$r^2 \geq 0.5$					
No. of precincts	55	78	43	44	26
No. of singletons[a]	23	31	16	16	11
No. of tagSNPs (Greedy)	81	110	71	66	41
No. of tagSNPs (FESTA)	75	105	67	61	38
No. of tagSNPs (FESTA double cover)	128	183	123	109	67
$r^2 \geq 0.8$					
No. of precincts	134	184	131	125	72
No. of singletons[a]	63	81	62	61	25
No. of tagSNPs (Greedy)	152	197	142	131	83
No. of tagSNPs (FESTA)	146	193	141	130	81
No. of tagSNPs (FESTA, double cover)	237	311	229	204	139

[a]Singleton refers to an SNP marker that is not in LD (pairwise LD measure r^2 greater than the threshold) with any other marker in the entire set. Such a marker, by definition, is one of the tagSNPs.

Table 4. Performance comparison of tagSNP sets selected by three different criteria in terms of coverage on denser SNP maps (CEU samples, with r^2 threshold of 0.5)

Region	ENr112	ENr131	ENr113	ENm010	ENm013
SNPs in dense map[a]	863	988	1061	539	708
SNPs in sparse map[b]	173	198	213	108	142
One-stage picking					
TagSNPs in dense map	75	105	67	61	38
Two-stage picking					
Max average r^2 b/tags and non-tags[c]	85	114	81	72	40
Min lowest r^2/tags and non-tags[d]	85	115	80	75	40
Min average r^2 among tags[e]	85	117	82	74	42
Other strategies					
Random picking[f]	103.2	137.7	91.4	71.0	52.0
Picket fence[g]	103	136	94	78	52
Use all sparse[h]	200	241	239	134	153

[a]Dense map means the densely typed SNP sets obtained in the ENCODE region from the HapMap website http://www.hapmap.org (CEU samples, release 16c, June 2005). All tagSNPs in this table were identified using our new algorithm.

[b]Sparse map means the SNP sets obtained by selecting the first SNP in every five consecutive SNPs in the dense maps.

[c–e]Total number of tagSNPs needed to cover the SNPs in the dense map using the tagSNPs identified using different criteria on the sparse map as seeds. Criterion 1, maximize the average r^2 between tagSNP and SNPs that it connected to; Criterion 2, minimize the average r^2 between tagSNP and SNPs that it connected to; Criterion 3, minimize the average r^2 among all tagSNPs.

[f]Total number of tagSNPs needed to cover the SNPs in the dense map using T random SNPs in the sparse map as seeds. T is the number of tagSNPs identified by our algorithm on the sparse map. The number is obtained by repeating this procedure 100 times and taking the average.

[g]Total number of tagSNPs needed to cover the SNPs in the dense map using T equally spaced SNPs (where we place equally spaced grid points along the interval and then select markers that are closest to these grid points) in the sparse map as seeds.

[h]Total number of tagSNPs needed to cover the SNPs in the dense map using all the SNPs in the sparse map as seeds.

fewer tagSNPs to cover the holes, compared with tagSNPs picked using a picket fence strategy (31.6% difference for r^2 threshold of 0.5 and 21.6% difference for r^2 threshold of 0.8) or picked at random (33.8% difference for r^2 threshold of 0.5 and 21.0% difference for r^2 threshold of 0.8).

DISCUSSION

In this manuscript, we developed an efficient computational framework for tagSNP selection using the pairwise r^2 criterion. Our algorithm is able to identify smaller tagSNP sets than the greedy approach (Carlson *et al*., 2004). Although the improvement is modest, our algorithm always outperforms the greedy approach in terms of the tagSNP size under exactly the same pairwise LD criterion. Using both chromosome-wide data and densely typed ENCODE region data from the HapMap Project, we illustrated the utility of our approach and showed savings increase in more densely typed regions and inside large LD 'blocks'. Computational time required by FESTA is quite reasonable and can be tailored to available computing resources as needed. Under the default setting, with r^2 threshold of 0.5, FESTA takes ∼1–15 min to run on the five ENCODE regions, and ∼120 min on entire Chromosome 2 (with r^2 threshold of 0.8, ∼0.1–1.5 min on the five ENCODE regions, and ∼24 min on Chromosome 2) using a 2.8 GHz Pentium class computer server. Another important advance is the ability of our method to identify multiple equivalent tagSNP sets and to use additional criteria to choose an optimal tagSNP set for typing. This feature offers flexibility in picking tagSNPs which is desirable when designing real association studies.

The key improvement of FESTA over the greedy approach is the 'precinct partitioning' step which enables the exhaustive search to be carried out very rapidly in most of the partitioned precincts. This is similar in spirit to the idea of 'partition-ligation' algorithm proposed by Niu *et al*. (2002) for haplotype inference.

Many of the existing tagSNP picking algorithms aim to capture haplotype diversity using the reduced set of markers (called haplotype tagging SNPs, htSNPs) such as BEST (Sebastiani *et al*., 2003). They work well when a small number of common haplotypes exist (typically true in the vicinity of a candidate gene) but these approaches often require the knowledge of complete haplotype phase and the boundary of the haplotype blocks. On the other hand, tagSNP selection using r^2 criteria does not require knowledge of block boundaries and can easily be applied to cover the whole chromosome. Recently, multiple-marker tagging strategies (Stram, 2005; P.I. de Bakker, 2005, http://www.broad.mit.edu/mpg/tagger) in which multiple tagSNPs can be used to represent each untagged SNPs have been proposed. While these methods further reduce the number of tagSNPs selected, this 'aggressive' approach may be sensitive to random genotyping failures.

Our approach is amenable to further computational improvements. For example, parallel programming could be used to search for tagSNPs in separate precincts, further speeding up the computation.

FESTA is freely available and can be downloaded at http://www.sph.umich.edu/csg/qin/FESTA

ACKNOWLEDGEMENTS

We are grateful to Drs Mike Boehnke, Randy Pruim and the three anonymous reviewers for critical comments on an early version of this manuscript. This work is partially supported by NIH RO1-HG002651-01 to G.A.

Conflict of Interest: none declared.

REFERENCES

Avi-Itzhak,H.I. *et al.* (2003) Selection of minimum subsets of single nucleotide polymorphisms to capture haplotype block diversity. *Pac. Symp. Biocomput.*, 466–477.

Cardon,L.R. and Abecasis,G.R. (2003) Using haplotype blocks to map human complex trait loci. *Trends Genet.*, **19**, 135–140.

Carlson,C.S. *et al.* (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analysis using linkage disequilibrium. *Am. J. Hum. Genet.*, **74**, 106–120.

Chapman,J.M. *et al.* (2003) Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum. Hered.*, **56**, 1831.

Collins,F.S. *et al.* (1997) Variations on a theme: cataloging human DNA sequence variation. *Science*, **278**, 1580–1581.

Cormen,T.H. *et al.* (2001) *Introduction to algorithms*. 2nd edition. MIT Press, Cambridge.

Daly,M.J. *et al.* (2001) High-resolution haplotype structure in the human genome. *Nat. Genet.*, **29**, 229–232.

Dawson,E. *et al.* (2002) A first generation slinkage disequilibrium map of human chromosome 22. *Nature*, **418**, 544–548.

Devlin,B. and Risch,N. (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*, **29**, 311–322.

Gabriel,S.B. *et al.* (2002) The structure of haplotype blocks in the human genome. *Science*, **296**, 2225–2229.

Goldstein,D.B. *et al.* (2003) Genome scans and candidate gene approaches in the study of common diseases and variable drug responses. *Trends Genet.*, **19**, 615–622.

Hampe,J. *et al.* (2003) Entropy-based SNP selection for genetic association studies. *Hum Genet.*, **114**, 36–43.

Hill,W.G. (1974) Estimation of linkage disequilibrium in randomly mating populations. *Heredity*, **33**, 229–239.

Hill,W.G. and Robertson,A. (1968) The effects of inbreeding at loci with heterozygote advantage. *Genetics*, **60**, 615–628.

Halldórsson,B.V. *et al.* (2004) Optimal haplotype block-free selection of tagging SNPs for genome-wide association studies. *Genome Res.*, **14**, 1633–1640.

Johnson,G.C. *et al.* (2001) Haplotype tagging for the identification of common disease genes. *Nat. Genet.*, **29**, 233–237.

Jeffreys,A.J. *et al.* (2001) Intensely punctuate meiotic recombination in the class II region of the major of histocompatibility complex. *Nat. Genet.*, **29**, 217–222.

Ke,X. and Cardon,L.R. (2003) Efficient selective screening of haplotype tag SNPs. *Bioinformatics*, **19**, 287–288.

Lin,Z. and Altman,R.B. (2004) Finding haplotype tagging SNPs by use of principal components analysis. *Am. J. Hum. Genet.*, **75**, 850–861.

Meng,Z. *et al.* (2003) Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. *Am. J. Hum. Genet.*, **73**, 115–130.

Niu,T. *et al.* (2002) Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am. J. Hum. Genet.*, **70**, 157–169.

Patil,N. *et al.* (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, **294**, 1719–1723.

Pritchard,J.K. and Przeworski,M. (2001) Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.*, **69**, 1–14.

Reich,D.E. *et al.* (2001) Linkage disequilibrium in the human genome. *Nature*, **411**, 199–204.

Rinaldo,A. *et al.* (2005) Characterization of multilocus linkage disequilibrium. *Genet. Epidemiol.*, **28**, 193–206.

Sachidanandam,R. *et al.* (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. International SNP Map Working Group (2001), *Nature*, **409**, 928–933.

Sebastiani,P. *et al.* (2003) Minimal haplotype tagging. *Proc. Natl Acad. Sci. USA*, **100**, 9900–9905.

Stram,D.O. *et al.* (2003) Choosing haplotype-tagging SNPs based on unphased genotype data using preliminary sample of unrelated subjects with an example from the multiethic cohort study. *Hum. Hered.*, **55**, 27–36.

Stram,D.O. (2005) Software for tag single nucleotide polymorphism selection. *Hum. Genomics*, **2**, 144–151.

The International HapMap Consortium (2003), The International HapMap Project. *Nature*, **426**, 789–796.

Zhang,K. *et al.* (2002) A dynamic programming algorithm for haplotype partitioning. *Proc. Natl Acad. Sci. USA*, **99**, 7335–7339.

Zhang,K. and Jin,L. (2003) HaploBlockFinder: haplotype block analysis. *Bioinformatics*, **19**, 1300–1301.